14321.31

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２０００年　６月２９日

出願番号
Application Number: 特願２０００－１９７４２１

出願人
Applicant (s): 株式会社エス・エス・アール
学校法人高知工科大学

CERTIFIED COPY OF
PRIORITY DOCUMENT

２０００年１２月　１日

特許庁長官
Commissioner,
Patent Office 及川耕造

出証番号　出証特２０００－３１００９６３

【書類名】　特許願

【整理番号】　PB06007

【提出日】　平成12年 6月29日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F 17/30

【発明の名称】　テキストマイニングにおける文書の特徴量抽出方法及びその装置

【請求項の数】　11

【発明者】

【住所又は居所】　高知県南国市蛍が丘１－１－１　株式会社エス・エス・アール内

【氏名】　吉岡　倍達

【発明者】

【住所又は居所】　高知県香美郡土佐山田町宮ノ口１８５　高知工科大学情報システム工学科内

【氏名】　ラック　ターウォンマット

【特許出願人】

【識別番号】　300044838

【氏名又は名称】　株式会社エス・エス・アール

【特許出願人】

【識別番号】　597154966

【氏名又は名称】　学校法人高知工科大学

【代理人】

【識別番号】　100077481

【弁理士】

【氏名又は名称】　谷　義一

【選任した代理人】

【識別番号】　100088915

【弁理士】

【氏名又は名称】　阿部　和夫

【選任した代理人】

【識別番号】　100106998

【弁理士】

【氏名又は名称】　橋本　傳一

【手数料の表示】

【予納台帳番号】　013424

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　0008215

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　テキストマイニングにおける文書の特徴量抽出方法及びその装置

【特許請求の範囲】

【請求項１】　文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出方法であって、前記単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算ステップと、前記単語－文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出ステップと、前記単語－文書行列を更新して前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分にする単語－文書行列更新ステップとを備えたことを特徴とするテキストマイニングにおける文書の特徴量抽出方法。

【請求項２】　前記コストは、前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分の二次コストとして定義されることを特徴とする請求項１に記載のテキストマイニングにおける文書の特徴量抽出方法。

【請求項３】　前記基底ベクトル計算ステップは、前記基底ベクトルの値を初期化する初期化ステップと、前記基底ベクトルの値を更新する基底ベクトル更新ステップと、前記基底ベクトルの値の変化度合いを求める変化度合い計算ステップと、前記基底ベクトルの値の変化度合いを用いて繰り返し処理を終了するかどうかを判別する判別ステップと、前記繰り返し処理の回数を数える計数ステップとを備えたことを特徴とする請求項１または２に記載のテキストマイニングにおける文書の特徴量抽出方法。

【請求項４】　前記基底ベクトル更新ステップは、前記基底ベクトルの現在値と、前記単語－文書行列と、前記基底ベクトルの更新度合いを制御する更新率とを用いて前記基底ベクトルを更新することを特徴とする請求項３に記載のテキ

ストマイニングにおける文書の特徴量抽出方法。

【請求項５】　前記特徴量の抽出に必要とされる全ての前記基底ベクトル及び前記正規化パラメータを既に取得している場合は、前記基底ベクトル計算ステップ及び前記特徴量抽出ステップにおける前記正規化パラメータの計算を省略し、前記特徴量抽出ステップは、既に取得している前記基底ベクトル及び前記正規化パラメータを用いて前記特徴量を抽出することを特徴とする請求項１～４のいずれか１項に記載のテキストマイニングにおける文書の特徴量抽出方法。

【請求項６】　文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出装置であって、前記単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算手段と、前記単語－文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出手段と、前記単語－文書行列を更新して前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分にする単語－文書行列更新手段とを備えたことを特徴とするテキストマイニングにおける文書の特徴量抽出装置。

【請求項７】　前記コストは、前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分の二次コストとして定義されることを特徴とする請求項６に記載のテキストマイニングにおける文書の特徴量抽出装置。

【請求項８】　前記基底ベクトル計算手段は、前記基底ベクトルの値を初期化する初期化手段と、前記基底ベクトルの値を更新する基底ベクトル更新手段と、前記基底ベクトルの値の変化度合いを求める変化度合い計算手段と、前記基底ベクトルの値の変化度合いを用いて繰り返し処理を終了するかどうかを判別する判別手段と、前記繰り返し処理の回数を数える計数手段とを備えたことを特徴とする請求項６または７に記載のテキストマイニングにおける文書の特徴量抽出装置。

【請求項９】　前記基底ベクトル更新手段は、前記基底ベクトルの現在値と、前記単語－文書行列と、前記基底ベクトルの更新度合いを制御する更新率とを用いて前記基底ベクトルを更新することを特徴とする請求項８に記載のテキストマイニングにおける文書の特徴量抽出装置。

【請求項１０】　前記特徴量の抽出に必要とされる全ての前記基底ベクトル及び前記正規化パラメータを既に取得している場合は、前記基底ベクトル計算手段及び前記特徴量抽出手段における前記正規化パラメータの計算を省略し、前記特徴量抽出手段は、既に取得している前記基底ベクトル及び前記正規化パラメータを用いて前記特徴量を抽出することを特徴とする請求項６～９のいずれか１項に記載のテキストマイニングにおける文書の特徴量抽出装置。

【請求項１１】　文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出装置において実行される特徴量抽出プログラム・プロダクトであって、前記単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算ステップと、前記単語－文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出ステップと、前記単語－文書行列を更新して前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分にする単語－文書行列更新ステップとを備えたことを特徴とする特徴量抽出プログラム・プロダクト。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、テキストマイニングにおける文書の特徴量抽出方法及びその装置に関し、より詳細には、特徴量を用いて文書および／またはウェブ検索、関連語検索、文書分類等の応用としてテキストマイニングを行う場合に、特徴量の空間において互いに関連した文書や単語が近接する特徴量を抽出するテキストマイニングにおける文書の特徴量抽出方法及びその装置に関する。

【０００２】

【従来の技術】

文章データを種々の観点から分析し、所望の知識や情報を取り出す技術であるテキストマイニングにおいて、文書の有効な特徴量抽出は文書および／またはウェブ検索、関連語検索、文書分類などを効率よく行うための重要な課題である。一般的な文書の特徴量抽出方法としては、「Automatic Text Processing」（Addison-Wesley社、１９８９年出版）の第３１３項で述べられているベクトル空間法（vector-space model）がよく用いられている。

【０００３】

ベクトル空間法では、文書の中で索引として選ばれた単語、即ち文書の内容を代表する索引語がｔ個ある場合、それぞれの索引語T_iにベクトルV_iを対応させ、ｔ次元のベクトル空間を定義する。このように定義されたベクトル空間を構成する全てのベクトルは、ｔ個の索引語に対応するｔ個のベクトルの線形結合として表現できる。このベクトル空間において、文書D_rを以下のように表現する。

【０００４】

【数１】

$$D=\sum_{i=1}^{t} x_{ir} V_i \qquad \text{式（１）}$$

【０００５】

式（１）において、V_iに作用するx_{ir}は文書D_rにおける索引語T_iに対する寄与分であり、文書D_rの特徴量を表す。特徴量とは、索引語の各文書における出現頻度を表す量である。ｔ×１（ｔ行１列）のベクトル［x_{r1}，x_{r2}，…，x_{rt}］′は文書D_rの特徴量ベクトルとなる。最も単純な場合としては、文書D_rにおいて索引語T_iが出現する場合には１とし、出現しない場合には０とする方法がとられる。より複雑な場合は、上記の文献の第２７９項から第２８０項までで述べられているように、文書D_rにおける索引語T_iの出現頻度（term frequency）tf_{ri}や、文書データベースに登録された全文書における索引語T_iを含む文書頻度df_iがx_{ir}の計算に利用される。

【０００６】

また、ｄ個の文書からなる文書の群に対しては、次のようなｔ×ｄの単語－文書行列Ｘが定義できる。

【０００７】

$$X=[x_1, x_2, \cdots, x_d]$$

ここで、ｔ次元のベクトル$x_j=[x_{j1}, x_{j2}, \cdots, x_{jt}]'$は文書$D_j$の特徴量ベクトルを表し、記号′は転置を示す。

【０００８】

図１は、文書データベースに登録された文書の一例を示す図である。また、図２は、図１に示された文書に出現する漢字の単語を索引語とした単語－文書行列の一例を示す図である。図２において、文書１～３の全てに出現している文字列「について教えて下さい」の中に含まれる「教」の文字は索引語の対象から外されている。図３は、ユーザから実際に入力される質問の一例を示す図である。この質問を図２の索引語を用いて表すと、図４に示す文書－単語行列で表現できる。

【０００９】

一般的に、ベクトル空間法を用いた場合、２つの文書D_rとD_sの類似度ｓｉｍ（D_r，D_s）は、以下のようになる。

【００１０】

【数２】

$$\mathrm{sim}(D_r, D_s)=\frac{\sum_{i=1}^{t} x_{ir}x_{is}}{\sqrt{\sum_{i=1}^{t} x_{ir}^2 \sum_{i=1}^{t} x_{is}^2}} \qquad \text{式（２）}$$

【００１１】

図３の質問の意味を基に、この質問と図１の各文書との類似度を判断した場合、図３の質問は図１の文書３に一番類似すると考えられる。しかし、図２及び図４のような特徴量ベクトルを用いると、図１における各文書と図３の質問の類似度は、それぞれ、ｓｉｍ（文書１，質問）＝０．５４７７、ｓｉｍ（文書２，質問）＝０．５４７７、ｓｉｍ（文書３，質問）＝０．５４７７となり、全ての文書に対して同じ類似度になってしまう。

【００１２】

このような問題点を解決する手法として「Journal of the American Society for Information Science」（１９９０年発行）の第４１巻第６号第３９１項から第４０７項までの記載において提案された、単語の共起に基づいた分析方法（Latent Semantic Analysis; ＬＳＡ）は、文書のもつ潜在的意味を抽出でき、かつ検索能率が圧倒的に優れている。ここにいう「単語の共起」とは、同一の文書／文に複数の単語が同時に出現することをいう。

【００１３】

ＬＳＡは、単語の共起の頻度を示す単語－文書行列を特異値分解（Singular Value Decomposition; ＳＶＤ）することにより、文書の潜在的意味構造を抽出するものである。得られた特徴量の空間において、互いに関連した文書や単語は近接するように構成される。「Behavior Research Methods, Instruments, & Computers」（１９９１年発行）の第２３巻第２号第２２９項から第２３６項までに掲載された論文では、ＬＳＡを使用した検索は、ベクトル空間法に比べ、３０％効率が良いという結果を報告している。以下、ＬＳＡについて具体的に説明する。

【００１４】

ＬＳＡでは、まずｔ×ｄの単語－文書行列Ｘを以下のように特異値分解する。

【００１５】

【数３】

$$X = T_0 S_0 D_0' \qquad \text{式（３）}$$

【００１６】

ここで、T_0はｔ×ｍの直交行列を表す。S_0はｍ個の特異値を対角要素とし、かつ対角要素以外はすべて０であるｍ×ｍの正方対角行列を表す。D'_0はｍ×ｄの直交行列を表す。また、０≦ｄ≦ｔとし、S_0の対角要素は値の大きい順に並んでいるものとする。

【００１７】

更に、ＬＳＡでは文書D_qのｔ×１の特徴量ベクトルx_qに対して次のような変換を行い、ｎ×１のＬＳＡ特徴量ベクトルy_qを計算する。

【００１８】

【数４】

$$y_q = S^{-1}T' x_q \qquad \text{式（４）}$$

【００１９】

ここで、ＳはＳ$_0$の対角要素の１番目からｎ番目までをとったｎ×ｎの正方対角行列、ＴはＴ$_0$の１列目からｎ列目まで抜き出したｔ×ｎの行列である。

【００２０】

例として、図２の単語－文書行列に対して特異値分解を行った結果を以下に示す。行列Ｔ$_0$、Ｓ$_0$、Ｄ$_0$はそれぞれ以下のようになる。

【００２１】

【数５】

$$T_0 = \begin{bmatrix} 0.1787 & -0.3162 & 0.3393 \\ 0.1787 & -0.3162 & 0.3393 \\ 0.1787 & -0.3162 & 0.3393 \\ 0.4314 & -0.3162 & -0.1405 \\ 0.4314 & -0.3162 & -0.1405 \\ 0.1787 & 0.3162 & 0.3393 \\ 0.1787 & 0.3162 & 0.3393 \\ 0.4314 & 0.3162 & -0.1405 \\ 0.4314 & 0.3162 & -0.1405 \\ 0.1787 & 0.3162 & 0.3393 \\ 0.2527 & 0.0000 & -0.4798 \end{bmatrix}$$

【００２２】

【数６】

$$S_0 = \begin{bmatrix} 2.7979 & 0 & 0 \\ 0 & 2.2361 & 0 \\ 0 & 0 & 1.4736 \end{bmatrix}$$

【００２３】

【数７】

$$D_0=\begin{bmatrix} 0.5000 & -0.7071 & 0.5000 \\ 0.5000 & 0.7071 & 0.5000 \\ 0.7071 & 0.0000 & -0.7071 \end{bmatrix}$$

【００２４】

ＬＳＡ特徴量ベクトルの次元ｔを２とし、図２の単語－文書行列の各特徴量ベクトルに対して式（４）を適用すると、文書１、２及び３のＬＳＡ特徴ベクトルはそれぞれ［０．５０００，－０．７０７１］′、［０．５０００，０．７０７１］′、［０．７０７１，０．００００］′となる。また、図４の特徴ベクトルに対して式（４）を適用すると、ユーザの質問のＬＳＡ特徴量ベクトルは［０．６５４２，０］′となる。

【００２５】

上記のように得られたＬＳＡ特徴量ベクトルに対して式（２）を適用し、図３の質問と図１に示した各文書との類似度を求めると、図１における各文書と図３質問の類似度は、それぞれ、ｓｉｍ（文書１，質問）＝０．５７７４、ｓｉｍ（文書２，質問）＝０．５７７４、ｓｉｍ（文書３，質問）＝１．０００となり、文書３が質問と一番類似するという結果が得られる。ネットワークを利用したヘルプシステムの応用などを想定する場合、図３の質問をしたユーザに対しては文書データベースに登録された文書３の回答文が返信されることになる。

【００２６】

特異値分解法は、一般的にThe Johns Hopkins University Press社が１９９６年に出版した「Matrix Computations」の第４５５項から第４５７項までの記載において提案されたアルゴリズムがよく用いられる。前記の「Journal of the American Society for Information Science」の論文によると、正方行列Ｓの行数（または列数）ｎの値は５０～１５０程度にすると良いとの記載がある。また、前記の「Behavior Research Methods, Instruments, & Computers」の論文において、ＬＳＡを行う前に特徴ベクトルの各要素を単に０または１の値をとると定義せずに、上記の出現頻度や文書頻度を用いて前処理するとより効果的であるという結果が報告されている。

【００２７】

【発明が解決しようとする課題】

しかし、上述の文献に提案されている特異値分解法のアルゴリズムでは、与えられた単語－文書行列から特徴量の空間を張る基底ベクトルを計算する過程において行列のバイダイアゴナリゼーション（bidiagonalization）のためにｔ×ｔの行列を利用するので、最低でも索引語数ｔの二乗t^2のオーダーのメモリ空間を必要とする。従って、従来の技術は、膨大な単語数又はデータ数を抱える文書データベースには適用できず、またデータ数の大小に関係なく行列の複雑な演算が必要であるという問題点があった。

【００２８】

本発明はこのような問題点に鑑みてなされたものであり、その目的とするところは、演算処理の容易化および当該演算処理に必要なメモリ容量の低減を図り、効率的に特徴量を抽出するテキストマイニングにおける文書の特徴量抽方法及びその装置を提供することにある。

【００２９】

【課題を解決するための手段】

本発明は、このような目的を達成するため、請求項１に記載の発明は、文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出方法であって、前記単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算ステップと、前記単語－文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出ステップと、前記単語－文書行列を更新して前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分にする単語－文書行列更新ステップとを備えたことを特徴とする。

【００３０】

また、請求項２に記載の発明は、請求項１に記載のテキストマイニングにおけ

る文書の特徴量抽出方法において、前記コストは、前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分の二次コストとして定義されることを特徴とする。

【００３１】

また、請求項３に記載の発明は、請求項１または２に記載のテキストマイニングにおける文書の特徴量抽出方法において、前記基底ベクトル計算ステップは、前記基底ベクトルの値を初期化する初期化ステップと、前記基底ベクトルの値を更新する基底ベクトル更新ステップと、前記基底ベクトルの値の変化度合いを求める変化度合い計算ステップと、前記基底ベクトルの値の変化度合いを用いて繰り返し処理を終了するかどうかを判別する判別ステップと、前記繰り返し処理の回数を数える計数ステップとを備えたことを特徴とする。

【００３２】

また、請求項４に記載の発明は、請求項３に記載のテキストマイニングにおける文書の特徴量抽出方法において、前記基底ベクトル更新ステップは、前記基底ベクトルの現在値と、前記単語－文書行列と、前記基底ベクトルの更新度合いを制御する更新率とを用いて前記基底ベクトルを更新することを特徴とする。

【００３３】

また、請求項５に記載の発明は、請求項１～４のいずれか１項に記載のテキストマイニングにおける文書の特徴量抽出方法において、前記特徴量の抽出に必要とされる全ての前記基底ベクトル及び前記正規化パラメータを既に取得している場合は、前記基底ベクトル計算ステップ及び前記特徴量抽出ステップにおける前記正規化パラメータの計算を省略し、前記特徴量抽出ステップは、既に取得している前記基底ベクトル及び前記正規化パラメータを用いて前記特徴量を抽出することを特徴とする。

【００３４】

また、請求項６に記載の発明は、テキストマイニングにおける文書の特徴量抽出装置において、文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出装置であって、前記単語－文書行列の各要素には前記索引語に対

する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算手段と、前記単語－文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出手段と、前記単語－文書行列を更新して前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分にする単語－文書行列更新手段とを備えたことを特徴とする。

【００３５】

また、請求項７に記載の発明は、請求項６に記載のテキストマイニングにおける文書の特徴量抽出装置において、前記コストは、前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分の二次コストとして定義されることを特徴とする。

【００３６】

また、請求項８に記載の発明は、請求項６または７に記載のテキストマイニングにおける文書の特徴量抽出装置において、前記基底ベクトル計算手段は、前記基底ベクトルの値を初期化する初期化手段と、前記基底ベクトルの値を更新する基底ベクトル更新手段と、前記基底ベクトルの値の変化度合いを求める変化度合い計算手段と、前記基底ベクトルの値の変化度合いを用いて繰り返し処理を終了するかどうかを判別する判別手段と、前記繰り返し処理の回数を数える計数手段とを備えたことを特徴とする。

【００３７】

また、請求項９に記載の発明は、請求項８に記載のテキストマイニングにおける文書の特徴量抽出装置において、前記基底ベクトル更新手段は、前記基底ベクトルの現在値と、前記単語－文書行列と、前記基底ベクトルの更新度合いを制御する更新率とを用いて前記基底ベクトルを更新することを特徴とする。

【００３８】

また、請求項１０に記載の発明は、請求項６～９のいずれか１項に記載のテキストマイニングにおける文書の特徴量抽出装置において、前記特徴量の抽出に必要とされる全ての前記基底ベクトル及び前記正規化パラメータを既に取得してい

る場合は、前記基底ベクトル計算手段及び前記特徴量抽出手段における前記正規化パラメータの計算を省略し、前記特徴量抽出手段は、既に取得している前記基底ベクトル及び前記正規化パラメータを用いて前記特徴量を抽出することを特徴とする。

【００３９】

更に、請求項１１に記載の発明は、文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて前記文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出装置において実行される特徴量抽出プログラム・プロダクトであって、前記単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する前記特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算ステップと、前記単語－文書行列及び前記基底ベクトルを用いて前記特徴量を正規化するためのパラメータを計算し、該パラメータに基き前記特徴量を抽出する特徴量抽出ステップと、前記単語－文書行列を更新して前記基底ベクトルを適用しない前記単語－文書行列と適用した前記単語－文書行列との差分にする単語－文書行列更新ステップとを備えたことを特徴とする。

【００４０】

本明細書によって開示される特徴量抽出装置は、以下の手段によって構成される。即ち、元の単語－文書行列と基底ベクトルを適用した単語－文書行列との差分の二次関数をコストとして定義し、そのコストに対して最急降下法を適用して基底ベクトルを計算する基底ベクトル計算手段と、単語－文書行列及び基底ベクトルを用いて、特徴量を正規化するためのパラメータを計算し、各文書に対して特徴量を抽出する特徴量抽出手段と、特徴量抽出手段の実行間で重複した特徴量を抽出しないように上記の差分で単語－文書行列を更新する単語－文書行列更新手段と、上記各手段の実行を制御する特徴量抽出制御手段とを備えていれば足りる。

【００４１】

基底ベクトル計算手段は、入力された単語－文書行列を基に計算を繰り返し、最終的に１つの基底ベクトルを算出する。繰り返しの処理は、各繰り返し処理間

で基底ベクトルの変化度合いが所定の基準値以下になったときに終了する。特徴量抽出手段は、入力された基底ベクトル及び単語－文書行列を基に、特徴量を正規化するためのパラメータを計算し、各文書に対して１つの特徴量を抽出する。単語－文書行列更新手段は、入力された基底ベクトルを基に、単語－文書行列を更新する。

【００４２】

特徴量抽出制御手段は、基底ベクトル計算手段、特徴量抽出手段、及び単語－文書行列更新手段を制御し、ユーザーにより定義された特徴量の数を満たすまで、各手段の実行を繰り返す。但し、基底ベクトル及び正規化パラメータが既に計算されている場合には、基底ベクトル計算手段の実行及び特徴量抽出手段における正規化パラメータの計算を省略される。そして、既に取得している基底ベクトル及び正規化パラメータを組み込んだ構成で特徴量抽出を行うことになる。

【００４３】

【発明の実施の形態】

図５は、本発明に係る特徴量抽出装置の一実施例を示す図である。図５に示すように、特徴量抽出制御手段２００は、単語－文書行列更新手段２１０と、基底ベクトル計算手段２２０と、特徴量抽出手段２３０とを備える。１００は単語－文書行列データファイル、３００は基底ベクトルデータファイル、４００は特徴量データファイル、４５０は正規化パラメータデータファイルである。単語－文書行列データファイル１００には、収集された文書データの単語－文書行列が記憶されている。単語－文書行列更新手段２１０は第１回目の繰り返し処理で単語－文書行列データファイル１００から単語－文書行列を読み込み、その単語－文書行列を更新せずに基底ベクトル計算手段２２０及び特徴量抽出手段２３０に渡す。

【００４４】

第２回目の繰り返し処理以降では、基底ベクトル計算手段２２０から渡された基底ベクトルを基に単語－文書行列を更新し、その結果を基底ベクトル計算手段２２０及び特徴量抽出手段２３０に渡す。基底ベクトル計算手段２２０は、単語－文書行列更新手段２１０から渡された単語－文書行列を基に繰り返し処理によ

り１つの基底ベクトルを計算する。そして、各繰り返し処理で基底ベクトルの変化度合いを監視し、変化度合いが所定の基準値以下になったときに繰り返しの処理を終了する。基底ベクトル計算手段２２０は、計算した基底ベクトルを基底ベクトルデータファイル３００に格納すると同時に、単語－文書行列更新手段２１０及び特徴量抽出手段２３０に渡す。特徴量抽出手段２３０は単語－文書行列更新手段２１０から渡された単語－文書行列及び基底ベクトル計算手段２２０から渡された基底ベクトルを基に各文書に対して１つの特徴量を抽出する。その結果を特徴量データファイル４００に格納すると同時に、それらの特徴量を正規化するためのパラメータを正規化パラメータデータファイル４５０に記録する。

【００４５】

単語－文書行列更新手段２１０，基底ベクトル計算手段２２０及び特徴量抽出手段２３０による、上述の実行を１回の繰り返しとする。繰り返し処理の回数を添字ｉで、ユーザーが指定した特徴量の数を添字ｎで示す。特徴量抽出制御手段２００では、ｉ＝ｎの条件を満たすまで、処理を一単位ずつ繰り返す。また、必要とされる全ての基底ベクトル及び正規化パラメータを既に取得しており、これらの値が既知の場合は、基底ベクトル計算手段２２０の実行及び特徴量抽出手段２３０における正規化パラメータの計算を省略し、既知の基底ベクトル及び正規化パラメータを組み込んだ単語－文書行列更新手段２１０及び特徴量抽出手段２３０のみで特徴量抽出制御手段２００を構成する。

【００４６】

図６は、本発明を実施するハードウェア構成の一例を示す図である。図６に示すように、特徴量抽出装置は、装置全体の制御を行う中央処理装置（Central Processor Unit；ＣＰＵ）１０と、プログラムが格納され又はプログラムの実行に必要な一時データ格納領域を提供するメモリ２０と、データを入力するためのキーボード３０と、表示画面を生成するディスプレイ４０とを備える。単語－文書行列データファイル１００、基底ベクトルデータファイル３００、特徴量データファイル４００、正規化パラメータデータファイル４５０及び特徴量抽出制御手段２００によって実行されるプログラムはメモリ２０に格納されている。

【００４７】

このような構成をとることにより、キーボード３０又はディスプレイ４０上の所定の位置指定するマウス等によりユーザーの指示を受けたＣＰＵ１０によって特徴量抽出が行われることとなる。なお、図５に示す例では、特徴量抽出制御手段２００はスタンドアロンの構成としているが、他のシステムに組み込んだ構成とすることも可能であることは言うまでもない。

【００４８】

図７は、単語－文書行列データファイルの構成図である。図７において、１０１－１，１０１－２，…，１０１－ｄはｄ個からなるｔ次元の単語－文書データに対応する。ここで、$X=[x_1, x_2, \cdots, x_d]$、$x_j=[x_{j1}, x_{j2}, \cdots, x_{jt}]'$を定義し、単語－文書データ１０１をｔ×ｄの行列Ｘで示す。

【００４９】

図８は、計算された基底ベクトルが格納された基底ベクトルデータファイルの構成図である。図８において、３０１－１，３０１－２，…，３０１－ｎはｎ個からなるｔ次元の基底ベクトルデータに対応する。ｉ番目の要素３０１－ｉは、図５におけるｉ回目の繰り返し処理における基底ベクトル計算手段２２０の出力値に対応する。以下の説明では、この要素をｔ×１の列ベクトル$w_i=[w_{i1}, w_{i2}, \cdots, w_{it}]'$で示す。

【００５０】

図９は、特徴量データファイルの構成図である。図９において、４０１－１，４０１－２，…，４０１－ｎはｎ個からなるｄ次元の特徴量データに対応する。ｉ番目の要素４０１－ｉは図５におけるｉ回目の繰り返し処理における特徴量抽出手段２３０による特徴量の出力値に対応する。この要素を１×ｄの行ベクトル$y_i=[y_{i1}, y_{i2}, \cdots, y_{id}]$で示す。

【００５１】

図１０は、正規化パラメータデータファイルの構成図である。図１０において、４５１－１，４５２－２，…，４５１－ｎはｎ個からなる正規化パラメータデータに対応する。ｉ番目の要素４５１－ｉは図５におけるｉ回目の繰り返し処理での特徴量抽出手段２３０による正規化パラメータの出力値に対応する。この要素をp_iで示す。

【００５２】

以上の諸定義を使用し、本実施形態に係る特徴量抽出の実現方式を詳細に説明する。単語－文書行列更新手段２１０では、ｉ＝１の場合、即ち繰り返し処理の１回目の実行に限り、Ｘを単語－文書行列データファイル１００から読み込み、何ら演算を行うことなくｔ×ｄの行列Ｅに格納する。従って、$E=[e_1, e_2, \cdots, e_d]$、$e_j=[e_{j1}, e_{j2}, \cdots, e_{jt}]'=[x_{j1}, x_{j2}, \cdots, x_{jt}]'$となる。前の繰り返し処理で抽出された特徴量が重複して抽出されないために、図５におけるｉ回目の繰り返しで下のようにＥをその現在値及び１つ前の繰り返し処理において計算された基底ベクトルを用いて更新し、その結果を基底ベクトル計算手段２２０に渡す。この処理によって格納される、Ｅのｉ番目の処理結果Ｅ（ｉ）は、式（５）のように表される。

【００５３】

【数８】

$$E(i)=\begin{cases}X, & \text{for } i=1\\ E(i-1)-w_{i-1}(w'_{i-1}E(i-1)), & \text{otherwise}\end{cases} \qquad \text{式（5）}$$

【００５４】

ここで、$E(i)=[e_1(i), e_2(i), \cdots, e_d(i)]$であり、Ｅ（ｉ）の各要素$e_j(i)$は$e_j(i)=[e_{j1}(i), e_{j2}(i), \cdots, e_{jt}(i)]'$で定義される。即ち、ｉ≧２の場合は、単語－文書行列は基底ベクトルを適用しない単語－文書行列から基底ベクトルを適用した単語－文書行列を引いた差分に更新される。

【００５５】

図１１は、基底ベクトル計算手段における基底ベクトルの計算の流れ図である。図１１におけるｋ回目の繰り返しでのw_iの値を$w_i(k)=[w_{i1}(k), w_{i2}(k), \cdots, w_{it}(k)]'$で示す。先ず、ステップＳ５００にて添字ｋを１で初期化する。続いてステップＳ５１０へ移行し、$w_i(1)$の各要素を－ＣからＣまでの間の任意の値で初期化する。ここで、Ｃの値は正の小さい数であり、例えばＣ＝０．０１としてもよい。ステップＳ５２０では、互いに関連した文書

や単語が近接する特徴量の空間を張る基底ベクトルを計算するため、式（６）に示す二次コストを設ける。

【００５６】

【数９】

$$\frac{1}{2d}\sum_{m=1}^{d}\sum_{l=1}^{t}(e_{lm}(i)-w_{li}\tilde{y}_{im})^2 \qquad \text{式（６）}$$

【００５７】

ここで、「単語が近接する」とは、特徴量の空間の中で複数の単語の位置が近接することを言い、「文書が近接する」とは、複数の文書の各々に含まれる単語の位置が特徴量の空間の中で近接することを言う。また、コストとは最小化したい対象を言い、本実施形態で定義されるコストは式（６）のように基底ベクトルを適用しない単語－文書行列と基底ベクトルを適用した単語－文書行列との差分の二次関数として定義される。ここで、

【００５８】

【外１】

$\tilde{y}_{im}$

【００５９】

は次のように定義される１×ｄのベクトル

【００６０】

【外２】

$\tilde{y}_{i}$

【００６１】

のｍ番目の要素である。

【００６２】

【数１０】

$$\tilde{y}_i=[\tilde{y}_{i1},\tilde{y}_{i2},.....,\tilde{y}_{id}]=w_i' E(i) \qquad \text{式（７）}$$

【００６３】

上記のコストに対して最急降下法を適用してw_iの値を式（８）のように更新する。

【００６４】

【数１１】

$$w_i(k+1) = w_i(k) + \frac{\mu_i(k)}{d}\,(E(i) - w_i(k)\,z_i(k))\,z_i(k)' \qquad \text{式（８）}$$

【００６５】

ここで、μ_i（ｋ）はｋ回目の繰り返しでの更新の度合いを制御する更新率で、ｋが１のときに正の小さい数で初期化し、例えばμ_i（１）＝０．１としてもよい。ｋが加算される度に徐々に値を減少させ、あるいはｋの値によらず一定値とすることも可能である。また、z_i（ｋ）は次のように定義される。

【００６６】

【数１２】

$$z_i(k) = w_i(k)'E(i) \qquad \text{式（９）}$$

【００６７】

ステップＳ５３０では次のようにw_iの変化度合いを示すδ_iを求める。

【００６８】

【数１３】

$$\delta_i(k) = \sqrt{\sum_{j=1}^{t}(w_{ji}(k+1) - w_{ji}(k))^2} \qquad \text{式（１０）}$$

【００６９】

ステップＳ５４０ではδ_i（ｋ）の値を基に処理を終了するかどうかを判別する。判別の結果、終了すると判断した場合はステップＳ５６０へ進み、そうでない場合はステップＳ５５０へ進む。ここで、図１１におけるβ_iは正の小さい数であり、例えば$\beta_i = 1 \times 10^{-6}$とすることができる。

【００７０】

ステップＳ５５０ではカウンタｋの値を１つ増やし、ステップＳ５２０に戻る。ステップＳ５６０ではw_iを基底ベクトルデータファイル３００にｉ番目のデータとして格納すると同時に、単語－文書行列更新手段２１０及び特徴量抽出手段２３０に渡す。特徴量抽出手段２３０では次のように特徴量y_i及び正規化パ

ラメータp_iを計算して、その結果それぞれを特徴量データファイル４００及び正規化パラメータデータファイル４５０にｉ番目のデータとして格納する。

【００７１】

【数１４】

$$y_i = \tilde{y}_i / p_i \ , \qquad \text{式(11)}$$

【００７２】

ここで、p_iは次のように定義される。

【００７３】

【数１５】

$$p_i = \sqrt{\sum_{j=1}^{d} \tilde{y}_{ij}^2} \qquad \text{式 (12)}$$

【００７４】

図１２は、本実施形態に係る特徴量抽出装置を用いた文書自動分類装置の一例を示す図である。図１２において、６０１は単語－文書行列計算手段、６０２は分類手段である。分類手段６０２は、１９９３年に発行された「Journal of Intelligent and Fuzzy Systems」の第１巻第１号第１項から第２５項で述べられている方法により行うことができる。

【００７５】

文書データベースに格納されている文書データは、文書自動分類装置に取り込まれる。文書自動分類装置では、単語－文書行列計算手段６０１で単語－文書行列の計算を行い、その結果を特徴量抽出制御手段２００に渡す。特徴量抽出制御手段２００は受け取った単語－文書行列から特徴量を抽出し、その結果を分類手段６０２に出力する。分類手段６０２では、入力された特徴量を基に分類の結果が出力される。

【００７６】

本発明の評価を、図１の文書や図３の質問のような文書データを含んだ、入試制度に関する実際の文書データにより特徴量抽出の評価を行った。本発明は、従来のＬＳＡを使用した場合と同じ性質の特徴量を抽出することが確認できた。

【００７７】

次に、使用されるメモリ空間のサイズに関して、単語数ｔが文書数ｄよりかなり大きい（ｔ≫ｄ）といった実際の場合において、従来のＬＳＡが少なくとも、t^2のオーダーを必要するのに対して、本発明は各々基底ベクトルの計算のために高々ｔ・ｄのオーダーのメモリサイズで足りる。また、従来の技術を実現するには、複雑な行列演算装置が必要であるが、本方式は四則演算程度を行う装置があれば容易に実現することができる。即ち、本発明によれば、ＬＳＡによる特徴量抽出と同等の効果を、より小さいメモリ空間、より簡単なプログラムにより得ることができる。また、この簡単なプログラムはＤＳＰ（Digital Signal Processor）におとすことができるため、特徴量抽出専用のチップを簡単に作成することが出来る。

【００７８】

以下、図１の文書及び図３の質問に対して本実施形態に係る特徴量抽出装置を実行した各手段の結果を示す。

【００７９】

Ａ．図１の文書

まず、図２の単語－文書行列をＸとする。

【００８０】

Ｉ．特徴量抽出制御手段２００における第１回目の繰り返し（ｉ＝１）

単語－文書行列更新手段２１０では式（５）より

【００８１】

【数１６】

$$E(1)=\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

【００８２】

を基底ベクトル計算手段２２０及び特徴量抽出手段２３０へ出力する。

【００８３】

基底ベクトル計算手段２２０では、w_1（１）を

[0.0100, −0.0100, 0.0100, −0.0100, 0.0100, −0.0100, 0.0100, −0.0100, 0.0100, −0.0100, 0.0100]′

で、μ_1を固定の値０．１で、β_1を1×10^{-6}で初期化し、以下のように図１１の計算を１３２回繰り返した後、基底ベクトルw_1＝

[0.1787, 0.1787, 0.1787, 0.4314, 0.4314, 0.1787, 0.1787, 0.4314, 0.4314, 0.1787, 0.2527]′

を基底ベクトルデータファイル３００に格納すると共に、特徴量抽出手段２３０、及び単語－文書行列更新手段２１０へ出力する。

【００８４】

＊基底ベクトル計算手段２２０における第１回目の繰り返し（ｋ＝１）

式（８）より、

w_1（２）＝

[0.0103, −0.0097, 0.0103, −0.0093, 0.0107, −0.0103, 0.0097, −0.0100, 0.0100, −0.0103, 0.0103]′

w_1（２）$-w_1$（１）$=10^{-3}\times$

[0.3332, 0.3334, 0.3332, 0.6668, 0.6666, −0.3332, −0.3334, 0.0001, −0.0001, −0.3332, 0.3332]′

δ_1（１）＝0.0013

＊基底ベクトル計算手段２２０における第２回目の繰り返し（ｋ＝２）

式（８）より、

w_1（３）＝

[0.0107, −0.0093, 0.0107, −0.0085, 0.0115, −0.0107, 0.0093, −0.0100, 0.0100, −0.0107, 0.0107]′

w_1（３）$-w_1$（２）$=10^{-3}\times$

[0.4110, 0.4112, 0.4110, 0.8001, 0.7998, −0.3665, −0.3668, 0.0224,

0.0221, −0.3665, 0.3887]′

δ_1（２）＝0.0015

中　略

＊基底ベクトル計算手段２２０における第１３２回目の繰り返し（ｋ＝１３２）

式（８）より、

w_1（１３３）＝

[0.1787, 0.1787, 0.1787, 0.4314, 0.4314, 0.1787, 0.1787, 0.4314, 0.4314, 0.1787, 0.2527]′

w_1（１３３）−w_1（１３２）＝10^{-6}×

[−0.3020, −0.3020, −0.3020, −0.3020, −0.3020, 0.3020, 0.3020, 0.3020, 0.3020, 0.3020, 0.0000]′

δ_1（１３２）＝9.5500×10^{-7}

特徴量抽出手段２３０では式（１１）及び式（１２）の演算を行い、

y_1＝[0.5000, 0.5000, 0.7071]

及び

p_1＝2.7979

をそれぞれ特徴量データファイル４００及び正規化パラメータデータファイル４５０へ出力する。

【００８５】

ＩＩ．特徴量抽出制御手段２００における第２回目の繰り返し（ｉ＝２）

単語－文書行列更新手段２１０では式（５）より

【００８６】

【数１７】

$$E(2)=\begin{bmatrix} 0.7500 & -0.2500 & -0.3536 \\ 0.7500 & -0.2500 & -0.3536 \\ 0.7500 & -0.2500 & -0.3536 \\ 0.3964 & -0.6036 & 0.1464 \\ 0.3964 & -0.6036 & 0.1464 \\ -0.2500 & 0.7500 & -0.3535 \\ -0.2500 & 0.7500 & -0.3535 \\ -0.6036 & 0.3965 & 0.1465 \\ -0.6036 & 0.3965 & 0.1465 \\ -0.2500 & 0.7500 & -0.3535 \\ -0.3536 & -0.3535 & 0.5000 \end{bmatrix}$$

【００８７】

を基底ベクトル計算手段２２０及び特徴量抽出手段２３０へ出力する。

【００８８】

基底ベクトル計算手段２２０では、w_2（１）を

[0.0100, −0.0100, 0.0100, −0.0100, 0.0100, −0.0100, 0.0100, −0.0100, 0.0100, −0.0100, 0.0100]′

で、μ_2を固定の値０．１で、β_2を1×10^{-6}で初期化し、図１１の計算を１１９回繰り返した後、基底ベクトルw_2＝

[0.3162, 0.3162, 0.3162, 0.3162, 0.3162, −0.3162, −0.3162, −0.3162, −0.3162, −0.3162, 0.0000]′

を基底ベクトルデータファイル３００に格納すると共に、特徴量抽出手段２３０、及び単語－文書行列更新手段２１０へ出力する。

【００８９】

＊基底ベクトル計算手段２２０における第１回目の繰り返し（ｋ＝１）

式（８）より、

$w_2(2)=$

[0.0102, −0.0098, 0.0102, −0.0096, 0.0104, −0.0105, 0.0095, −0.0103, 0.0097, −0.0105, 0.0102]′

$w_2(2)-w_2(1)=10^{-3}\times$

[0.2154, 0.2156, 0.2154, 0.3822, 0.3821, −0.4511, −0.4513, −0.2844

, －0.2846, －0.4511, 0.1666]′

δ_2（１）＝0.0011

＊基底ベクトル計算手段２２０における第２回目の繰り返し（ｋ＝２）

式（８）より、

w_2（３）＝

[0.0105, －0.0095, 0.0105, －0.0092, 0.0108, －0.0110, 0.0090, －0.0106, 0.0094, －0.0110, 0.0103]′

w_2（３）－w_2（２）＝10^{-3}×

[0.2624, 0.2626, 0.2624, 0.4413, 0.4411, －0.5152, －0.5154, －0.3364, －0.3366, －0.5152, 0.1786]′

δ_2（２）＝0.0013

中　略

＊基底ベクトル計算手段２２０における第１１９回目の繰り返し（ｋ＝１１９）

式（８）より、

w_2（１２０）＝

[0.3162, 0.3162, 0.3162, 0.3162, 0.3162, －0.3162, －0.3162, －0.3162, －0.3162, 0.0000]′

w_2（１２０）－w_2（１１９）＝10^{-6}×

[0.3327, 0.3333, 0.3327, －0.1375, －0.1381, 0.3332, 0.3326, －0.1377, －0.1383, 0.3332, －0.4712]′

δ_2（１１９）＝9.8141×10^{-7}

－特徴量抽出手段２３０では式（１１）及び式（１２）の演算を行い、

y_2＝[0.7071, －0.7071, －0.0000]

及び

p_2＝2.2361

をそれぞれ特徴量データファイル４００及び正規化パラメータデータファイル４

５０へ出力する。

【００９０】

上記の結果から図１における文書１，２，３の特徴量ベクトルはそれぞれ［0.5000，0.7071］′、［0.5000，－0.7071］′、［0.7071，－0.0000］′となる。これらは、従来例の説明において示された各文書のＬＳＡの特徴量と比較すると、第二番目の要素の符号が逆になっているが同一の絶対値をとる。従って、式（２）の類似度の計算に関してＬＳＡの特徴量と同じ性質を持つ。

【００９１】

Ｂ．図３の質問

ここでは、図１の文書の特徴量抽出の際に基底ベクトルデータファイル３００に格納された基底ベクトル及び正規化パラメータデータファイル４５０に格納された正規化パラメータを用いるので、基底ベクトル計算手段２２０の実行及び特徴量抽出手段における正規化パラメータの計算を省略する。図３の質問をＸとする。

【００９２】

Ｉ．特徴量抽出手段２００における第１回目の繰り返し（ｉ＝１）

単語－文書行列更新手段２１０では、式（５）より

【００９３】

【数１８】

$$E(1)=\begin{bmatrix}0\\0\\1\\1\\1\\1\\1\\0\\1\\0\\0\end{bmatrix}$$

【００９４】

を特徴量抽出手段２３０へ出力する。

【００９５】

特徴量抽出手段２３０では、図１の文書の特徴量抽出の際に得られた特徴量ベクトルw_1及び正規化パラメータp_1を用いて式（１１）及び式（１２）の演算を行い

$$y_1 = [0.6542]$$

を特徴量データファイル４００へ出力する。

【００９６】

ＩＩ．特徴量抽出制御手段２００における第２回目の繰り返し（ｉ＝２）

単語－文書行列更新手段２１０では、図１に示す文書の特徴量抽出を行う際に得られた特徴量ベクトルw_1を用いて、式（５）より

【００９７】

【数１９】

$$E(2) = \begin{bmatrix} -0.3271, \\ -0.3271, \\ 0.6729, \\ 0.2103, \\ 0.2103, \\ 0.6729, \\ 0.6729, \\ -0.7897, \\ 0.2103, \\ -0.3271, \\ -0.4626 \end{bmatrix}$$

【００９８】

を特徴量抽出手段２３０へ出力する。

【００９９】

特徴量抽出手段２３０では図１の文書の特徴量抽出の際に得られた特徴量ベクトルw_2及び正規化パラメータp_2を用いて、式（１１）及び式（１２）の演算を行い、

$$y_2 = [-0.0000]$$

を特徴量データファイル４００へ出力する。

【０１００】

上記の結果から図３の質問の特徴量ベクトルは［0.6542, －0.0000］′となり、従来例の説明で示した値と比較すると２番目の要素は同一の絶対値をとる。

【０１０１】

【発明の効果】

以上説明したように、本発明によれば、文書の内容を代表する索引語に対応するベクトルからなる単語－文書行列を用いて文書の特徴量を抽出するテキストマイニングにおける文書の特徴量抽出方法であって、単語－文書行列の各要素には前記索引語に対する寄与分が作用し、コストを最小化する最急降下法に基いて互いに関連した文書および単語が近接する特徴量の空間を張る基底ベクトルを計算する基底ベクトル計算ステップと、単語－文書行列及び基底ベクトルを用いて特徴量を正規化するためのパラメータを計算し、パラメータに基き特徴量を抽出する特徴量抽出ステップと、単語－文書行列を更新して基底ベクトルを適用しない単語－文書行列と適用した単語－文書行列との差分にする単語－文書行列更新ステップとを備えたので、テキストマイニングにおける文書の特徴量抽出に関し、ＬＳＡを実行可能な装置よりも小さいメモリ空間でＬＳＡと同じ性質を持つ特徴量を抽出することができる。また、ＬＳＡと同じ性質を持つ特徴量を抽出するための専用ソフトウェアやハードウェアを容易に作成することが可能となる。

【図面の簡単な説明】

【図１】

文書データベースに登録された文書の一例を示す図である。

【図２】

図１に示された文書に出現する漢字の単語を索引語とした単語－文書行列の一例を示す図である。

【図３】

ユーザから実際に入力される質問の一例を示す図である。

【図４】

図３から得られた単語－文書行列を示す図である。

【図５】

本発明に係る特徴量抽出装置の一実施例を示す図である。

【図６】

本発明を実施するハードウェア構成の一例を示す図である。

【図７】

単語－文書行列データファイルの構成図である。

【図８】

計算された基底ベクトルが格納された基底ベクトルデータファイルの構成図である。

【図９】

特徴量データファイルの構成図である。

【図１０】

正規化パラメータデータファイルの構成図である。

【図１１】

基底ベクトル計算手段における基底ベクトルの計算の流れ図である。

【図１２】

本発明の一実施形態に係る特徴量抽出装置を用いた文書自動分類装置の一例を示す図である。

【符号の説明】

１０　ＣＰＵ

２０　メモリ

３０　キーボード

４０　ディスプレイ

１００　単語－文書行列データファイル

１０１－１、１０１－２、１０１－３、１０１－ｄ　単語－文書データ

２００　特徴量抽出制御手段

２１０　単語－文書行列更新手段

２２０　基底ベクトル計算手段

２３０　特徴量抽出手段

３００　基底ベクトルデータファイル

３０１－１、３０１－２、３０１－３、３０１－ｎ　基底ベクトルデータ

４００　特徴量データファイル

４０１－１、４０１－２、４０１－３、４０１－ｎ　特徴量データ

４５０　正規化パラメータデータファイル

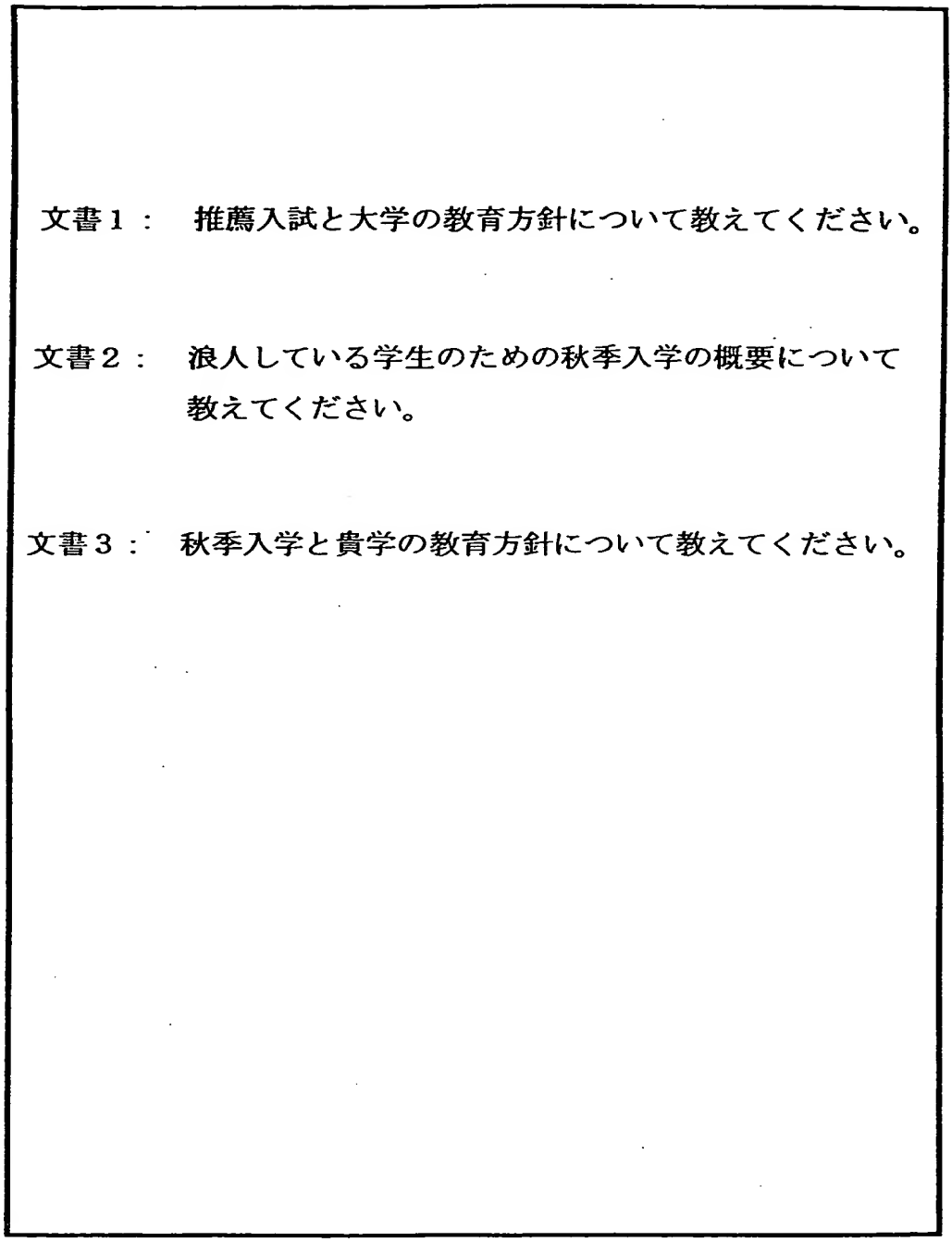
４５１－１、４５１－２、４５１－３、４５１－ｎ　正規化パラメータデータ

６０１　単語－文書行列計算手段

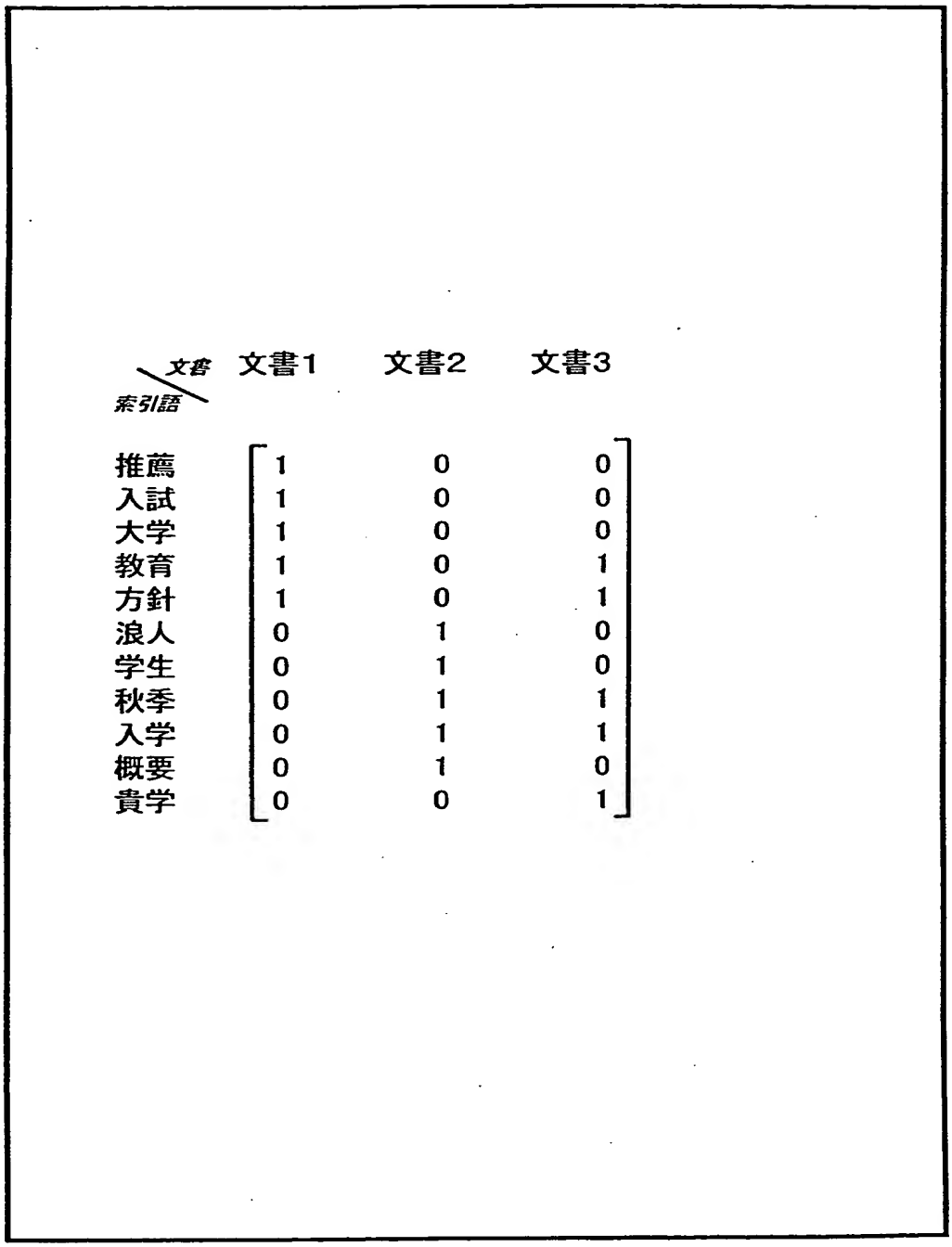
６０２　分類手段

【書類名】　　図面

【図１】


文書１：　推薦入試と大学の教育方針について教えてください。
文書２：　浪人している学生のための秋季入学の概要について教えてください。
文書３：　秋季入学と貴学の教育方針について教えてください。

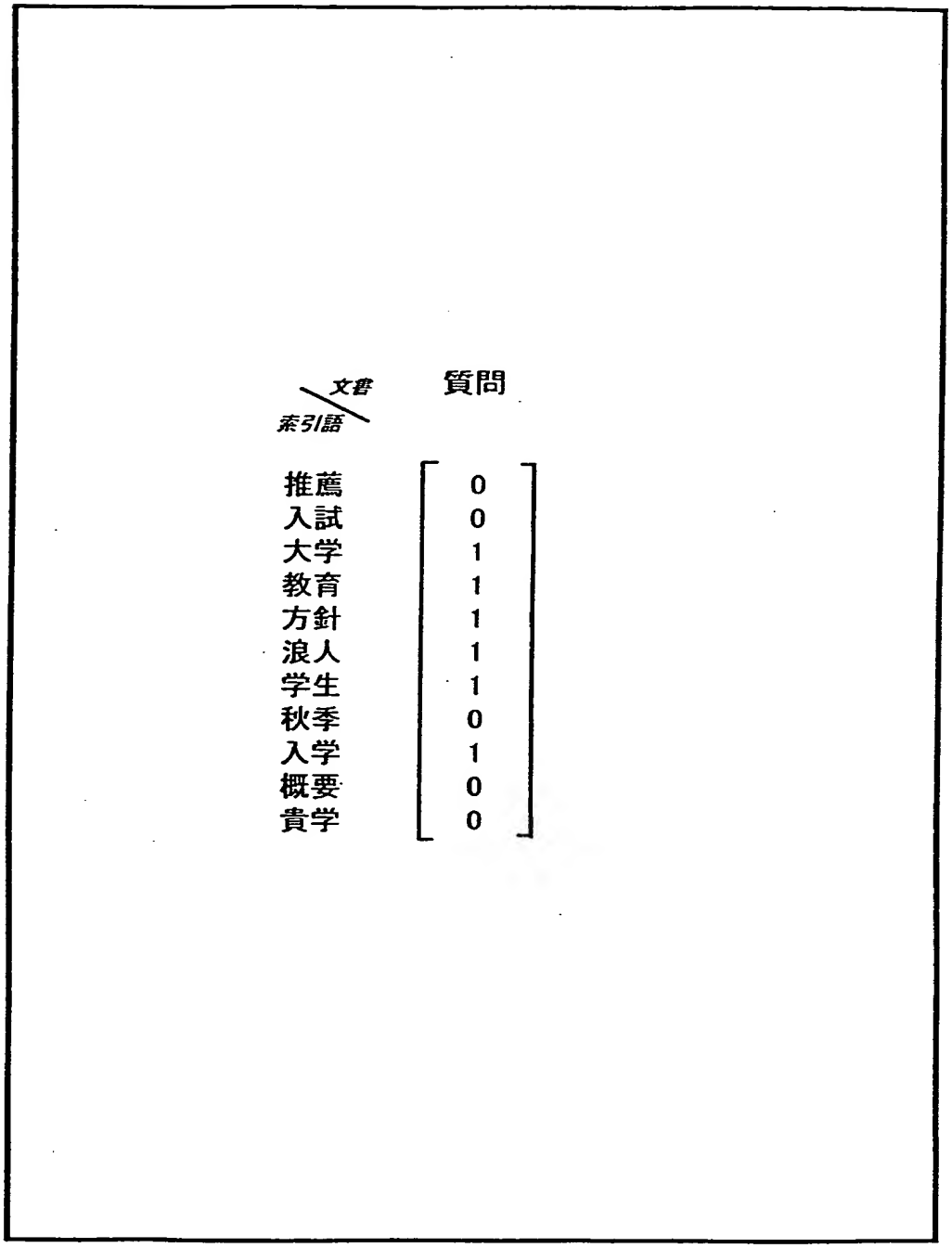
【図２】


文書
索引語
文書1
文書2
文書3
推薦 1 0 0
入試 1 0 0
大学 1 0 0
教育 1 0 1
方針 1 0 1
浪人 0 1 0
学生 0 1 0
秋季 0 1 1
入学 0 1 1
概要 0 1 0
貴学 0 0 1

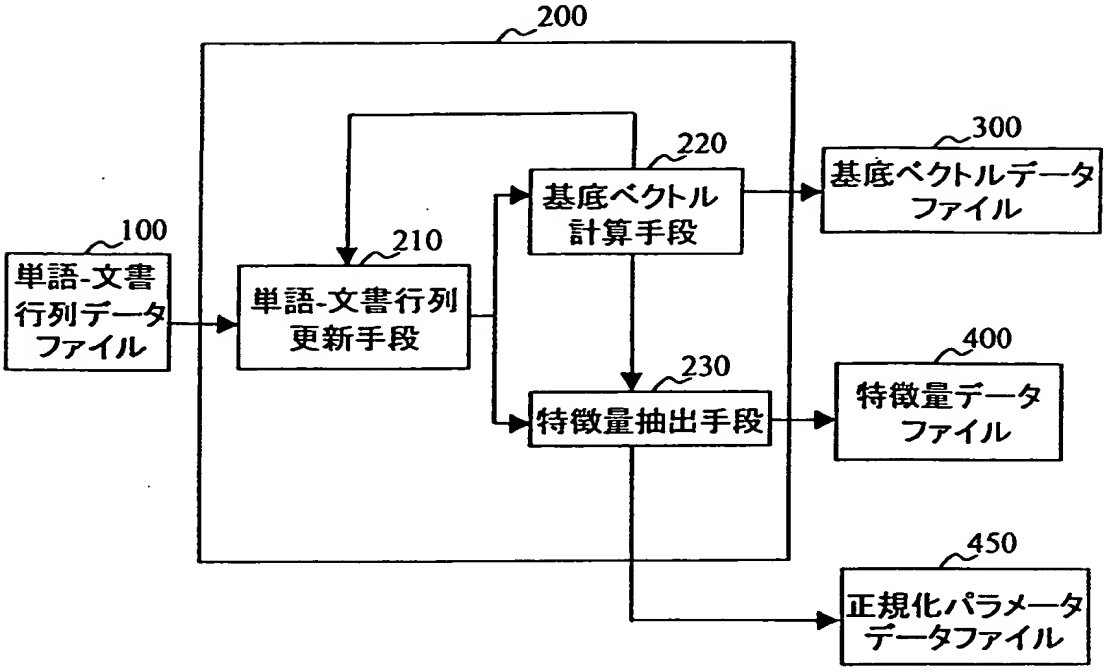
【図３】

質問：　浪人の学生を対象した入学の制度及び大学の教育方針について教えてください。

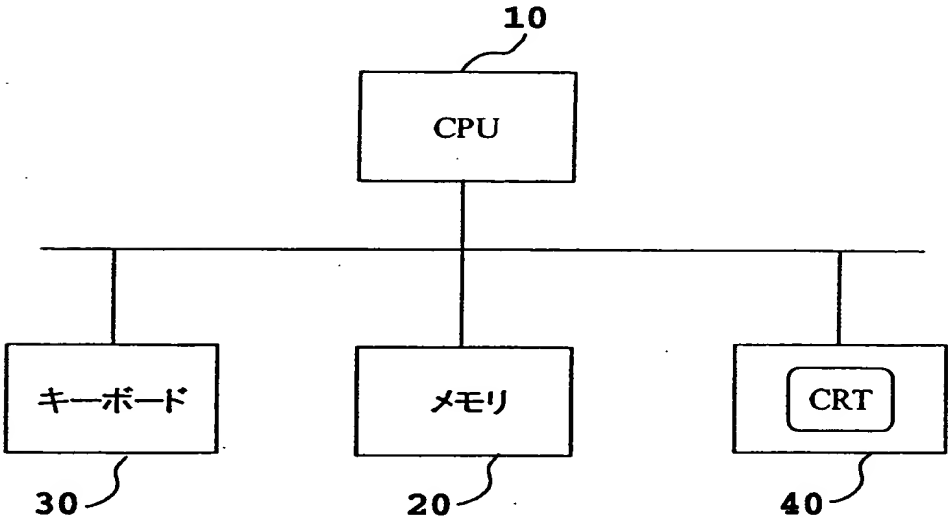
【図４】



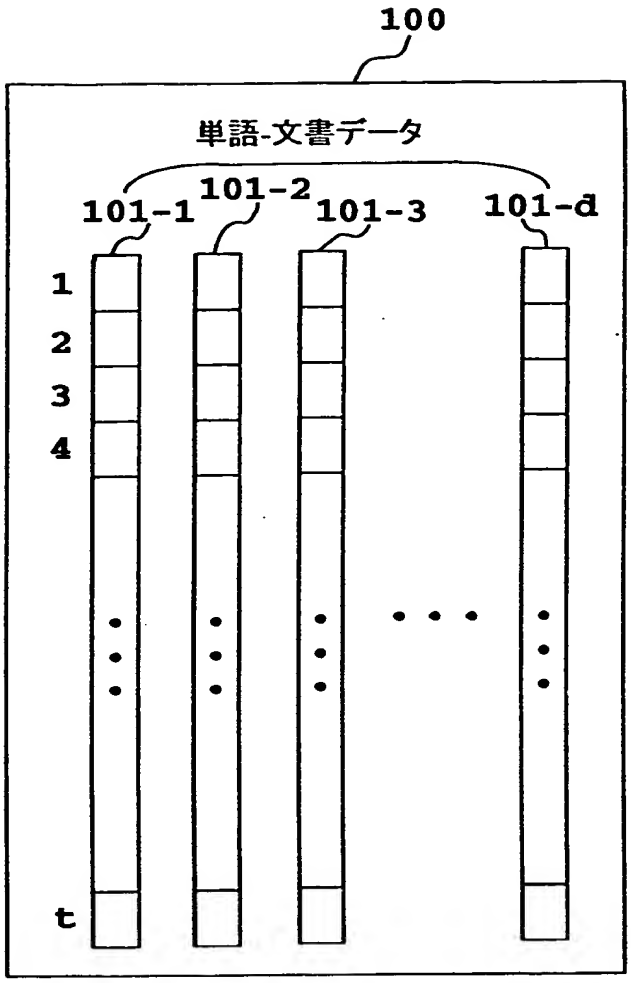
【図５】


200
100
単語-文書
行列データ
ファイル
210
単語-文書行列
更新手段
220
基底ベクトル
計算手段
300
基底ベクトルデータ
ファイル
230
特徴量抽出手段
400
特徴量データ
ファイル
450
正規化パラメータ
データファイル

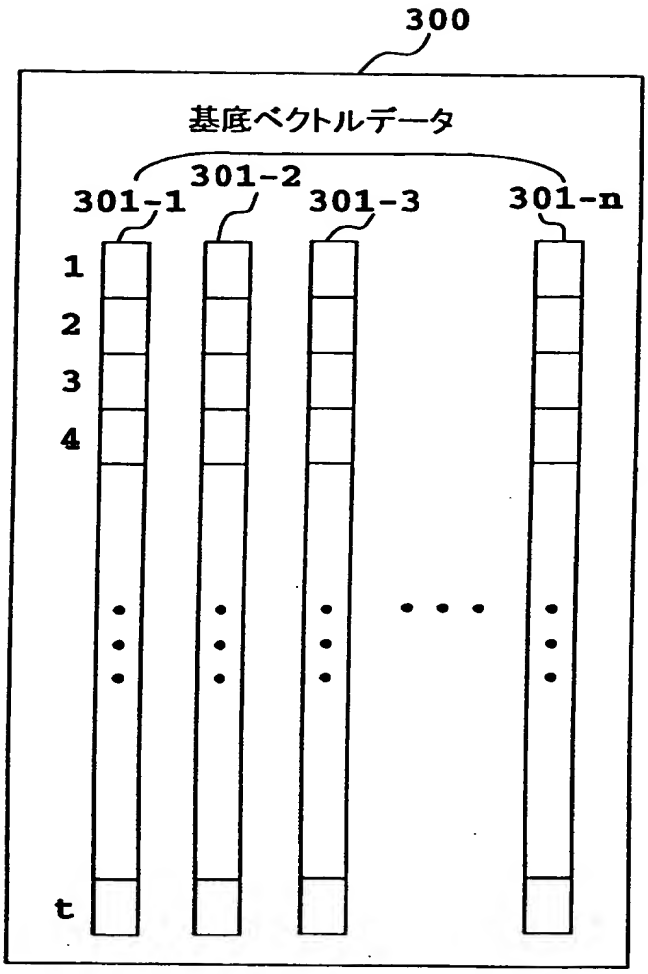
【図６】


10
CPU
キーボード
メモリ
CRT
30
20
40

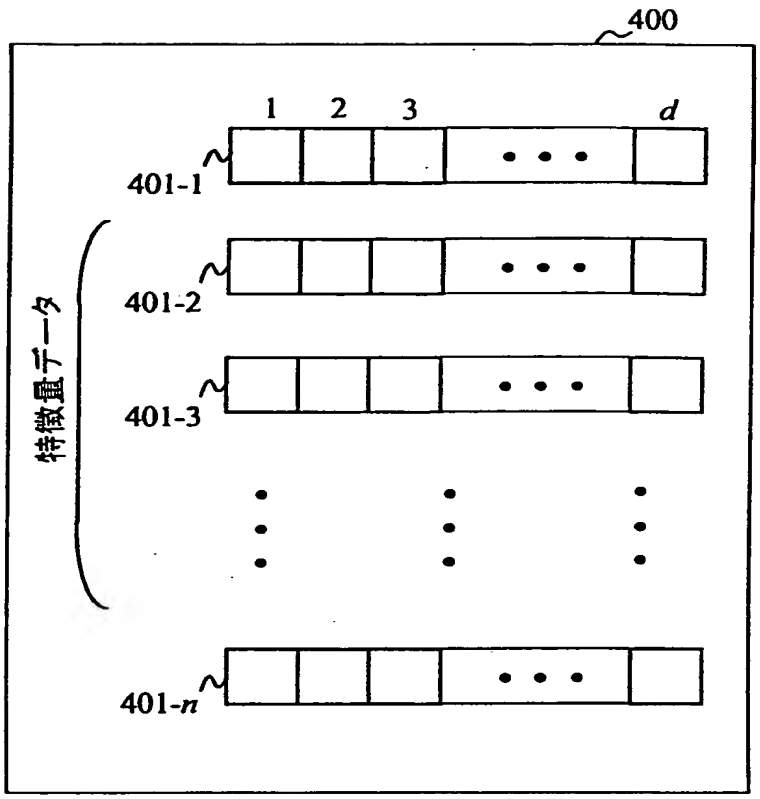
【図７】



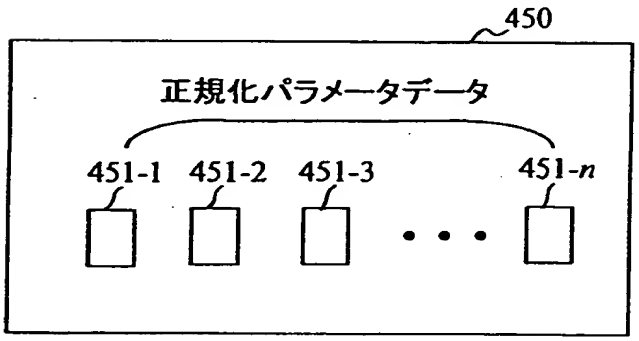
【図８】



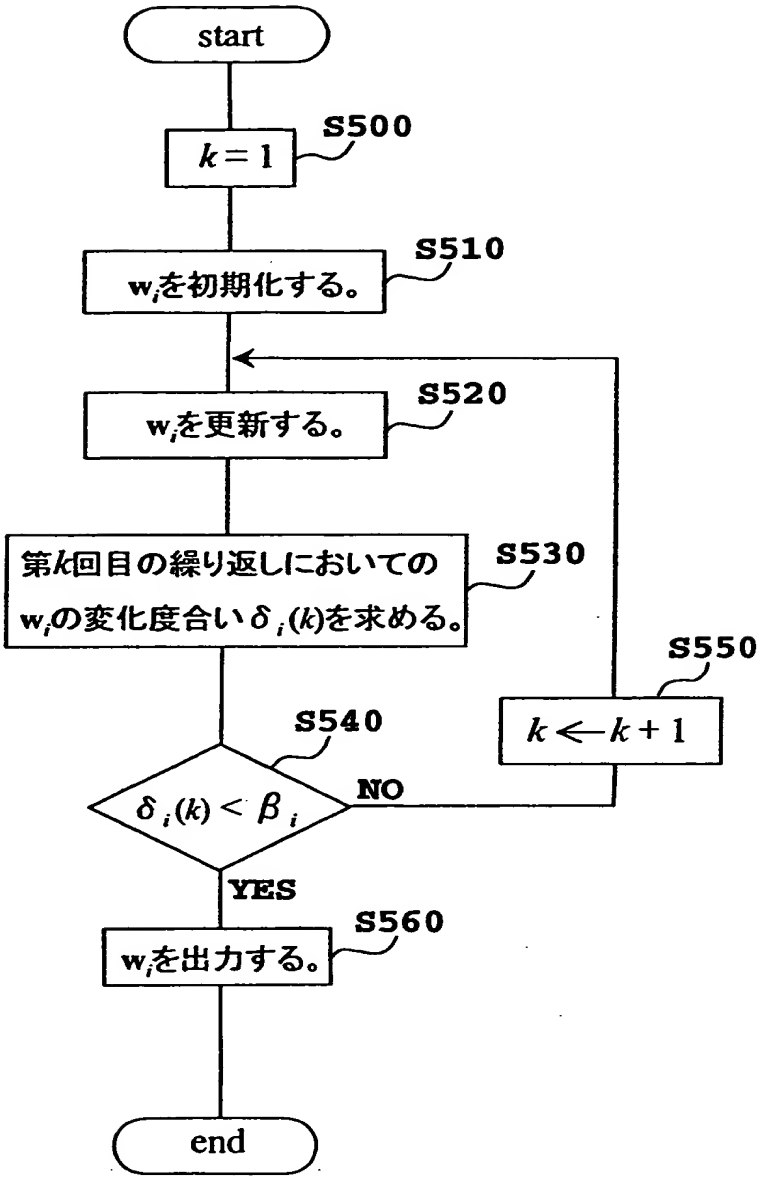
【図９】


400
1
2
3
d
401-1
401-2
401-3
401-n
特徴量データ

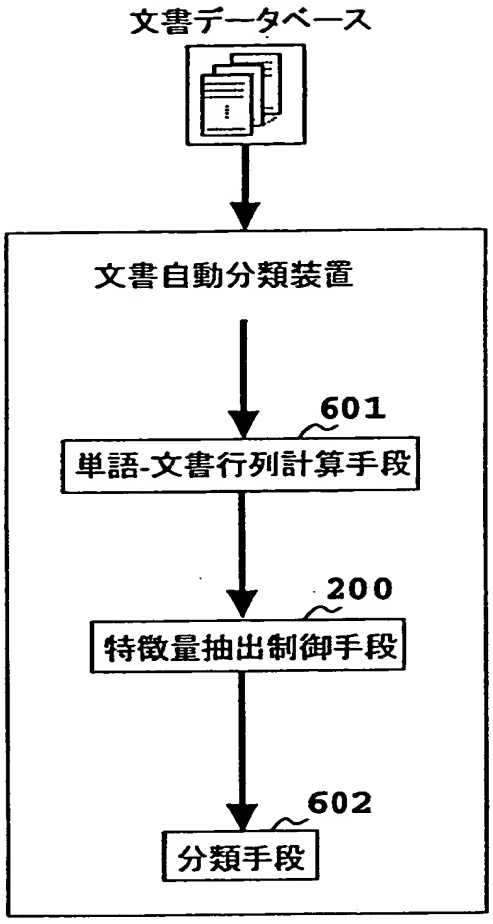
【図１０】


450
正規化パラメータデータ
451-1
451-2
451-3
451-n

【図１１】


start
S500
$k=1$
S510
w_iを初期化する。
S520
w_iを更新する。
第k回目の繰り返しにおいての
w_iの変化度合い$\delta_i(k)$を求める。
S530
S550
$k \leftarrow k+1$
S540
$\delta_i(k) < \beta_i$
NO
YES
S560
w_iを出力する。
end

【図１２】



【書類名】　　要約書

【要約】

【課題】　テキストマイニングにおける文書の特徴量抽出に関し、ＬＳＡを実行可能な装置よりも小さいメモリ空間及びより簡単なプログラム及び装置で、ＬＳＡと同じ性質を持つ特徴量を抽出する方法及びその装置を提供する。

【解決手段】　単語－文書行列更新手段２１０により更新された単語－文書行列、及びその行列を基に基底ベクトル計算手段２２０で計算された、有効な特徴量の空間を張る基底ベクトルを基に、特徴量抽出手段２３０で各文書の特徴量を抽出し、特徴量がユーザーにより与えられた所定の数を満たすまで各手段の実行を繰り返す。

【選択図】　　図５

出　願　人　履　歴　情　報

識別番号　　〔３００００４４８３８〕

１．変更年月日　　２０００年　６月　１日
［変更理由］　新規登録
住　所　　高知県南国市蛍が丘１－１－１　南国オフィスパークセンター３Ｆ
氏　名　　株式会社エス・エス・アール

出　願　人　履　歴　情　報

識別番号　　〔５９７１５４９６６〕

１．変更年月日　　１９９７年１１月　５日
［変更理由］　新規登録
住　所　　高知県香美郡土佐山田町宮ノ口１８５番地
氏　名　　学校法人高知工科大学